Comparison of oligonucleotide microarray and RNA-SEQ technologies in the context of gene expression analysis

Nicolas Sierro, Florian Martin, Carine Poussin, Julia Hoeng, Nikolai V. Ivanov✉
Philip Morris R&D, Neuchatel, Switzerland

Motivation and Objectives

For more than a decade, the microarray technology has been widely and extensively used to profile gene expression in various study types. Transcript abundance measurement is currently revolutionized by RNA-seq technology. Indeed, RNA-seq enables the sequencing of the whole transcriptome (sequence-centric data) while only predefined transcripts/genes can be measured on arrays (gene-centric data). In addition, the nature of RNA-seq data renders the analysis more flexible for addressing biological questions going from transcript/gene (differential) quantification to transcript structure (splice variants) identification (qualitative analysis) to cite a few. This latest application requires the use of additional specific arrays (e.g. exon), which have some limitations. The purpose of our work here was to compare both affymetrix GeneTitan array and Illumina HiSeq-2000 sequencing technologies in the context of gene expression analysis.

Methods

For this study, mRNA samples from lung tissue of ApoE mice exposed to conventional cigarette smoke (CS) or fresh air (Sham) for 3 and 6 months, or from ApoE mice exposed to CS for 3 months and then exposed to fresh air for 3 months (Cessation) were hybridized on Affymetrix MG-HT430PM GenTitan array or sequenced on Illumina HiSeq-2000 (2x100bp paired-end run, ~30 to 100 million paired reads per sample). Quality control analysis was performed for each data type accordingly. RNA-seq data were cleaned using the fastx toolkit. Briefly, 3'-ends of reads were trimmed with a quality threshold of 20, and filtered to retain only reads of at least 50 bases, at least 90% of which with a base-calling quality of 20 or more. Quantification at the gene level was performed using RSEM and VOOM-LIMMA to accommodate for mean-variance trend or Cufflinks followed by Cuffdiff for differential count analysis. Array data were pre-processed using RMA and differential expression was carried out using LIMMA.

Results and Discussion

Comparing gene expression abundances measured by both technologies revealed a significant correlation for highly expressed genes, while this correlation significantly decreased for low expressed abundance genes. The Illumina HiSeq-2000 showed a wider range of expression values than the Affymetrix GeneTitan array for low expressed genes. This result indicates that the Illumina HiSeq-2000 sequencing technology has a higher sensitivity to detect low expressed genes. When comparing differential expression (treatment vs control samples), a higher fold change magnitude was observed in the volcano plot of data generated with the Illumina HiSeq-2000. The magnitude of the significance of the observed fold changes was similar between both platforms. However, in-more-depth analysis varying False Discovery Rate and Fold Change thresholds showed that Illumina HiSeq-2000 is more sensitive to detect differentially expressed genes. Further investigations will be undertaken to understand if these low expressed transcripts/genes detected by RNA-seq reveal new biological functions or not compared to those identified with the array technology. Overall, this study shows that RNA-seq is a very powerful technology since it is more sensitive to detect low and differential expressed transcripts/genes compared to the Affymetrix GeneTitan technology.

Acknowledgements

The project is funded by PMI.